Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations

Katarzyna Bryc[a,*], Wlodek Bryc[b], Jack W. Silverstein[c]

[a] *Department of Genetics, Harvard Medical School, Boston, MA 02115, USA*
[b] *Department of Mathematical Sciences, University of Cincinnati, PO Box 210025, Cincinnati, OH 45221–0025, USA*
[c] *Department of Mathematics, Box 8205, North Carolina State University, Raleigh, NC 27695-8205, USA*

Abstract

We present a mathematical model, and the corresponding mathematical analysis, that justifies and quantifies the use of principal component analysis of biallelic genetic marker data for a set of individuals to detect the number of subpopulations represented in the data. We indicate that the power of the technique relies more on the number of individuals genotyped than on the number of markers.

Keywords:
Principal Components Analysis, Eigenanalysis, Population Structure, Eigenvalues, Number of Subpopulations

1. Introduction

Principal component analysis (PCA) has been a powerful and efficient method for analyzing large datasets in population genetics since its early applications by Cavalli-Sforza and others (Menozzi et al., 1978; Cavalli-Sforza et al., 1993, 1994). In particular, PCA of single nucleotide polymorphism (SNP) genotype data can be used to illuminate population structure (Nelson et al., 2008), provide

*Corresponding author, *Phone:* 617-432-1101
Email addresses: kbryc@genetics.med.harvard.edu (Katarzyna Bryc), Wlodzimierz.Bryc@uc.edu (Wlodek Bryc), jack@ncsu.edu (Jack W. Silverstein)

insights into human history and admixture (Novembre et al., 2008; McVean, 2009), and help to estimate the number of distinct subpopulations within a sample (Patterson et al., 2006). PCA can also be applied to correct for population structure within a sample of individuals, to prevent spurious results in conducting medical genetic studies such as genome-wide association studies (GWAS), which seek to find genes underlying diseases or traits (Price et al., 2006; Zhu et al., 2002).

In this paper, we provide additional mathematical confirmation for the use of PCA in estimating the number of subpopulations within a sample. In a related result (Patterson et al., 2006, Theorem 3) that motivated this research, the authors analyze the theoretical centered covariance matrix for a single marker as the number of individuals increases without bound. Here we analyze a mathematically more complicated object: the sample covariance matrix based on multiple markers. In current practice, the sample covariance matrix is often centered, and the data rows are often further normalized. In contrast to previous work, our results describe behavior of the eigenvalues of the sample covariance matrix *without* centering or normalization, taking into account both the number of individuals and the number of markers. The raw unprocessed covariance matrix is more amenable to mathematical analysis, and the singular values of such raw data exhibit quantifiable properties that can be used directly to determine the number of populations in the data in an almost deterministic fashion, at least when the number of individuals in the study is sufficiently large.

We show that for large data sets of individuals from K well-differentiated subpopulations, with overwhelming probability the un-centered sample covariance matrix has K large eigenvalues. (The technical meaning of "well differentiated subpopulations" is that matrix $\mathbf{Q}$, which we later define in equation (4.8) from the pairwise moments of pairwise site spectra, is non-singular.) These large eigenvalues, which indicate the presence of population structure, are greater by

a factor proportional to the number of individuals than the remaining smaller eigenvalues. The large eigenvalues arise from the mixed moments of the pairwise site frequency spectra, while the small eigenvalues are attributed to random differences between the individuals in the sample. In practice, with finite populations, we can detect only the eigenvalues that are well separated from zero where the cutoff described in equation (2.1) is beyond the boundary of the range of the many smaller eigenvalues, which we will refer to as the "bulk" of the eigenvalues.

We note that the *eigenvectors* of a sample covariance matrix are also interesting, but notoriously difficult to analyze mathematically, so this paper is devoted solely to understanding the eigenvalues.

2. Methods

In setting up the mathematical model, we begin as in Patterson et al. (2006). We consider unrelated diploid individuals with independent biallelic markers. We assume that the data for our biallelic markers are recorded in a large $M \times N$ rectangular array $\mathbf{C}$ with rows labeled by individuals and columns labeled by polymorphic markers. The entries $C_{i,j}$ are the number of variant alleles for marker j, individual i, that take values $0, 1$ or 2. We assume that we have data for M individuals from K populations, and that we have M_r individuals from the subpopulation labeled r so that $M = M_1 + M_2 + \cdots + M_K$. Often, neither K nor $M_1, \ldots, M_K$ are known, so we may wish to estimate the value of K, the number of subpopulations in the data. If the population sampling information were known, namely, that individual i is from subpopulation r, the genotype probabilities for marker j, $\mathbb{P}(C_{i,j} = 0, 1, 2)$ would be given by the expected frequencies in population r, as in equations (4.1)-(4.3).

In what follows we deviate from the method of Patterson et al., since for mathematical analysis it is inconvenient to rely on data-dependent statistics (of

mean and variance) to center or to normalize the entries of the array. Instead we work directly with the eigenvalues of the uncentered sample covariance matrix $\mathbf{CC}'$. This is a symmetric square matrix of size M, the number of individuals. We consider the eigenvalues of $\mathbf{CC}'$ which we write in decreasing order $\Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_M$.

In this paper we formally derive the mathematical proof of an estimator for the number of subpopulations based on the magnitude of these eigenvalues. We derive an estimate for the number of subpopulations, K, as the number of eigenvalues larger than the threshold of

$$t' = \frac{1+F}{2}\left(\sqrt{M}+\sqrt{N}\right)^2 = N\frac{1+F}{2}\left(1+\sqrt{M/N}\right)^2 \tag{2.1}$$

Equivalently, for the scaled matrix $\mathbf{X}$ we later define in equation (4.10) we can use the more intuitive threshold:

$$t = \frac{1+F}{2} \tag{2.2}$$

which does not depend on M, N. The parameter F used here, as defined by equation (4.4), takes values between 0 and 1. In practice, F captures departures from Hardy-Weinberg equilibrium.

Hence begins our main result, that depending on the value of F, the threshold cutoff t for determining the number of large eigenvalues corresponding to population structure, is between 0.5 and 1. If there are K subpopulations present in the data, then as N and M increase without bound (and are subject to certain technical conditions), with overwhelming probability the smallest $M-K$ eigenvalues of $\mathbf{CC}'$ are smaller than t' from (2.1). Furthermore, the consecutive largest K eigenvalues are typically much larger than the order of $4\lambda_j MN$, where $\lambda_j > 0$ is the corresponding eigenvalue of the deterministic (hidden) matrix with entries given by (4.8) which we describe below. Since we write all eigenvalues in decreasing order, our ability to correctly estimate the value of K depends on

the magnitude of the smallest eigenvalue λ_K in comparison to the number of individuals M.

Here, in evaluating whether an eigenvalue corresponds to population structure, we are effectively comparing a constant between 1/2 and 1 (depending on the value of F), to a number larger than $\lambda_K M$. Of course, $\lambda_K M$ can be made arbitrarily large by increasing the number of individuals M, making it possible to easily resolve eigenvalues corresponding to population structure separate from the bulk. From our mathematical analysis, we show that the number of populations is estimated essentially without error when $M\lambda_K$ is larger than 1. In view of this strong separation, the eigenvalues of $\mathbf{CC}'$, can be safely used in exploratory data analysis without need for a formal statistical test to assess significance of structure when M is large enough.

A check for the appropriateness of the cutoff is provided by the histogram of the eigenvalues – the K largest eigenvalues should be separated from the remaining eigenvalues, or the bulk. Under the model of clean population substructure, the remaining eigenvalues should cluster together into a fairly solid group, as these eigenvalues correspond to random differences among individuals. Ideally, one expects the shape of the bulk to be a single-mode semi-elliptical mass located with sharp boundaries like the Marchenko-Pastur law (Bai and Silverstein, 2010, Chapter 3). After normalization (4.10), the distribution of the bulk should be located to the left of $(1 + F)/2$.

The accuracy of this estimator of K depends on the theoretical smallest population eigenvalue:

$$L = \frac{4MN\lambda_K}{\left(\sqrt{M} + \sqrt{N}\right)^2} \tag{2.3}$$

That is, it depends on the smallest of the theoretical population eigenvalues of (3.1), for the rescaled matrix $\mathbf{X}_N$ from equation (4.10). This theoretical value indicates whether there is likely to be power to detect the full population sub-

structure present in the data, if, for example, $L > 0.5$. Indeed, our simulations confirm that our estimate of population structure works very well whenever L is larger than 0.5 (when $F = 0$), or L is larger than 1 (when $F = 1$). In practice, population parameter λ_K (the smallest eigenvalue of matrix $\mathbf{Q}$, see equation (4.8)) and thereby L, is a hidden parameter that depends on the theoretical hidden population moments and on the unknown relative proportions $c_1, c_2, \ldots, c_K$ of the subpopulations present in the data, and cannot be obtained for non-simulated datasets.

2.1. Robust to violations of assumptions

In our analysis we explore possible violations of our key assumptions, namely, independence among markers and stochastic independence of individuals drawn from a population.

Simulations indicate that linkage disequilibrium (LD), the non-random correlation of nearby markers that violates our independence assumption, does not strongly affect our ability to detect population structure. In our view, the thinning of markers, such as via the LD-pruning implemented in *PLINK*, is a simple yet robust technique that addresses linkage disequilibrium violations of independence assumptions, that works without the need for corrections described in (Patterson et al., 2006; Shriner, 2012). Our formulas show that no matter what the value of λ_K is, one can safely reduce the value of N by thinning, or removing nearby markers that are highly correlated, without a significant loss of accuracy. For example, to compensate for thinning which would reduce the number of markers $N = 500M$ to $N' = 100M$, it is enough to increase M by about 20%. To compensate for the potential loss of accuracy due to thinning down to $N' = 10M$, one needs to increase M by about 73%. Furthermore, if λ_K is far enough from zero so that L is much larger than 0.5, then thinning will have no effect even if the number of individuals M is kept unchanged. Our analysis shows that the

number of subpopulations K determined by this procedure is quite insensitive to the number of markers N, as long as this number is well above the number of individuals M.

A possible violation of the assumption of stochastic independence of individuals is non-random mating, which results in departures from Hardy-Weinberg equilibrium (HWE). We find that that departures from HWE do not significantly reduce the power of PCA for detecting population substructure. In fact, to compensate for $F = 1$ instead of $F = 0$ it is enough to increase the number of individuals M in the study by a factor of 2; however, no such corrections are needed if the smallest eigenvalue λ_K is separated from zero well enough so that $L > 0.5$.

Lastly, our application to real genotype data from the International HapMap Project (HapMap) suggests that cryptic relationships between the individuals in the data may affect the applicability of our method to a larger degree than LD. Under a simple substructure scenario, the bulk of eigenvalues should have a unimodal elliptical shape similar to the Marchenko-Pastur distribution, easily distinguished from large eigenvalues corresponding to substructure. However, as we demonstrate in Figure 2, individuals may exhibit cryptic relatedness, or other unknown non-random relationships, which result in changes to the distribution of the bulk, making it difficult to infer the correct cutoff for substructure. We find that pruning for LD does not seem to improve the fit of the bulk to Marchenko-Pastur distribution. Instead, exclusion of related individuals improves fit of the bulk; hence, we suggest that it is necessary to remove related individuals from the sample to improve the resolution of true substructure.

2.2. Summary

Overall, our mathematical analysis confirms empirical evidence that PCA is a robust technique for learning about population substructure of a dataset.

Contrary to current practice, based on the mathematical theory presented in the following sections we recommend using PCA directly on the data matrix $\mathbf{C}$ without centering or renormalization. With sufficient sample size, there should be strong separation between the large eigenvalues corresponding to population structure and the remaining bulk of the distribution. We illustrate a proof of principle of our approach through simulations and application to human genotype data from world-wide populations.

3. Results and Discussion

In this section we illustrate the power of our mathematical findings for inference of population structure in genetic data. We begin with the simulations for a "simple model" where all the hidden parameters can be computed. This allows us to analyze sensitivity of the technique to the precise value of L in (2.3). In particular, since we are able to compute the hidden parameter L, we can then see how well our predictions match theory, and how well powered we are to detect the known substructure. Then we consider an intermediate stage – we use simulations from (Gao et al., 2011) where the true demography is known and each individual is a member of one of the populations, but for which the hidden population parameters are not known. The datasets cover several different demographic models, with different population split times, trees, and migration rates. For more details on each of the models, see reference (Gao et al., 2011). Finally, we apply the theory to human genotype data from world-wide populations, where we discuss additional challenges due to linkage disequilibrium and cryptic relatedness, and where the value of L is not available.

3.1. Simulations for a simple model

We show that the theoretical approximations to the largest eigenvalues work very well when all the assumptions of mathematical analysis are satisfied. We gen-

erate simulations based on a simple model in which we make several assumptions that are unrealistic, but allow us to compute important mathematical parameters to explore the performance of our method. We assume that the site frequency spectra are known for each population. We also know how many individuals came from each subpopulation, and that the populations are independent. The latter corresponds to a scenario where all populations diverged and stopped interacting in the distant past. Though unrealistic, such a simplistic model has the advantage that all relevant quantities that enter mathematical analysis can be computed. In particular, this model allows us to study the effects of choosing a small enough number of individuals M and analyze how the failure rate for the estimator depends on the value of L, see Table 1.

In our simulations we use unequal subpopulation samples sizes, drawn with proportions $c_1 = 1/6$, $c_2 = 1/3$, $c_3 = 1/2$. The theoretical population proportion $p_r(j)$ at each SNP location for each population was selected from the same site frequency spectrum $\varphi(x) = 0.5/\sqrt{x}$. We selected $p_1(j), p_2(j), p_3(j)$ independently at each location j which corresponds to product joint site frequency spectrum $\varphi(x, y, z) = \varphi(x)\varphi(y)\varphi(z)$ for our $K = 3$ simulated populations. We then simulated independent individual genotypes for the j-th marker of a member of the r-th population by choosing independent binomial values (with 2 trials) with probability of success $p_r(j)$. We can evaluate how well the mathematical description matches the simulated data because we can explicitly compute the theoretical matrix of moments (4.5-4.6) and hidden matrix $\mathbf{Q}$ defined by (4.8):

$$[m_{r,s}] = \begin{bmatrix} 1/5 & 1/9 & 1/9 \\ 1/9 & 1/5 & 1/9 \\ 1/9 & 1/9 & 1/5 \end{bmatrix}, \ \mathbf{Q} = [\sqrt{c_r c_s}\, m_{r,s}] = \begin{bmatrix} 0.0333 & 0.0262 & 0.0321 \\ 0.0262 & 0.0667 & 0.0454 \\ 0.0321 & 0.0454 & 0.1000 \end{bmatrix}$$

The eigenvalues of the above matrix $\mathbf{Q}$ are $[\lambda_1, \lambda_2, \lambda_3] = [0.1467, 0.0355, 0.0178]$. The theoretical prediction for the observed largest eigenvalues of the normalized

sample covariance matrix (4.10) are then

$$\Lambda_j \approx \frac{4MN\lambda_j}{\left(\sqrt{M}+\sqrt{N}\right)^2}. \tag{3.1}$$

The actual observed eigenvalues will not match exactly these predictions; the purpose of the simulations is to illustrate how far the empirical values for finite M, N differ from the values predicted by theory in the limit as M, N tend to infinity. For example, formula (3.1) with $M = 1200, N = 25000$ gives the following values: $(47.4, 11.5, 5.7)$. In a simulation, we obtained the following eigenvalues for the normalized matrix (4.10):

$$(\mathbf{\Lambda_1}, \mathbf{\Lambda_2}, \mathbf{\Lambda_3}, \Lambda_4, \Lambda_5, \dots) = (\mathbf{48.2}, \mathbf{11.5}, \mathbf{5.8}, 0.27, 0.26, \dots)$$

We see that the threshold of 0.5 separates clearly the $K = 3$ largest eigenvalues, set in boldface, from the bulk.

We remark that there are two sources of error: the approximation $\mathbf{B}_N \approx \mathbf{Q}$ that appears in Lemma 2 and then the approximation due to randomness within each population that is still present in (4.12) for finite numbers of SNPs, N. It is therefore encouraging to see that the predicted values for the eigenvalues match well with the empirical eigenvalues in the simulations for realistic values of $M =$ 1,200, $N =$ 25,000 as well was for much smaller values of M, see Table 1, or even for $M = 12$. When $M = 24$ and $N = 100$, we get $L = 0.77$ and we are successful in determining correct value of $K = 3$ the vast majority of the time: in 100,000 simulations, K is underestimated in only a minuscule 0.005 % of the runs and never overestimated. Reducing further the number of individuals to $M = 12$, leads to $L = 0.471$ and in this case, as expected, the rate at which our estimate of K fails increases. But the decrease in accuracy is not dramatic, and an underestimate of $\hat{K} = 2$ occurs only in about 6.5% of runs. For reasonably large M and N, our power is quite high, and the false positive error rate was

strictly zero for all scenarios, since an overestimate of K did not occur in any of the replicates under any scenario.

Table 1 False negative probability as a function of L (based on 100,000 simulations)

M	N	L	$\mathbb{P}(\hat{K} < 3)$
individuals	*SNPs*		*False negative rate*
1200	25000	5.747	0
48	100	1.926	0
24	100	0.770	0.00005
12	100	0.471	0.065
12	50	0.385	0.53
6	100	0.276	0.87

3.2. Simulated genetic data under various demographic scenarios

Next we applied our method to simulated substructured datasets generated by coalescent simulations under various demographic scenarios from (Gao et al., 2011). This dataset has $N = 100$ markers sampled from subpopulations with constant population sizes of 50 individuals, and with varying $M = 50 \times K$, $K = 1, \ldots, 5$.

In these simulations, $\mathbf{Q}$ is not known. But we can estimate L by using the smallest eigenvalue of the empirical approximation to $\mathbf{Q}$ based on formula (4.16). The observed accuracy under each scenario, shown in Tables 2–4, is in line with our results from simulations listed in Table 1.

Model-based approaches such as those evaluated in (Gao et al., 2011) are likely to outperform PCA detection for such small sample sizes, since the true substructure corresponds to that found in the underlying *STRUCTURE*-like model (Falush et al., 2003). However, using our method, we have no false positives in

any of these sets of simulations. From these simulations we find that the error rates are not affected by using approximate $\mathbf{Q}$ in the calculations instead of exact $\mathbf{Q}$ when we approximate L from the population data. (The values $\hat{L}$ of the approximated L varied considerably in the simulated 50 runs for a model, but $\hat{L} > 0.5$ was associated with the correct value of $\hat{K}$ in each case.)

Table 2 False negative (error) rates of estimates of K for 50 simulated data sets under model *Split*

True K	2	3	4	5
$\mathbb{P}(\hat{K} < K)$	0.0	0.14	0.80	0.98

Table 3 False negative (error) rates of estimates of K for 50 simulated data sets under model *Inbred*

True K	2	3	4	5
$\mathbb{P}(\hat{K} < K)$	0.0	0.16	0.72	1.0

Table 4 False negative (error) rates of estimates of K for 50 simulated data sets under model *Mig*

True K	2	3	4	5
$\mathbb{P}(\hat{K} < K)$	0.0	0.02	0.10	0.46

3.3. Application to human population genotype data

We next examined the distribution of eigenvalues for a dataset of human genotype data. The International HapMap Project was designed to create a catalog of human genetic variation to find genes that affect health, disease, and individual responses to medications and environmental factors. We use a genome-wide SNP dataset made publicly available through this project as HapMap 3 (HapMap 3, release 3, human genome build 36) which contains genotypes of

individuals from 11 human populations, comprised of genotype data collected using two platforms: the Illumina Human1M and the Affymetrix SNP 6.0 arrays. These populations and datasets have been extensively studied previously (see `http://hapmap.ncbi.nlm.nih.gov/publications.html.en` for a list of publications).

Unlike simulated data, the true substructure of the complete set of populations is unknown. We therefore report the performance of our theoretical analysis on the subset of well defined subpopulations, which are believed to have clear substructure: the Yoruba, of Ibadan, Nigeria (YRI), European Americans from Utah (CEU), and Chinese from Denver (CEU).

After extracting the CEU, CHB, and YRI individuals, we processed the data through PLINK (Purcell et al., 2007) with filters `--filter-founders --geno 0` to remove SNPs with any missing data and exclude offspring of trios, and further exclude non-autosomal markers. The final dataset of $M = 297$ individuals and $N = 736750$ markers was used for the analysis of the eigenvalues.

As expected from mathematical theory and from the choice of very distinct subpopulations, the eigenvalues of matric $\mathbf{X}$ split into the bulk in Figure 1 that lies below the cutoff of 0.5, and three large eigenvalues $\Lambda_1 = 102.0$, $\Lambda_2 = 14.55$, and $\Lambda_3 = 7.37$ that exceed the cutoff of 0.5 and give an estimate of $\hat{K} = 3$, which matches our prediction for these three populations. The histogram seems to show some possible eigenvalues separated from the bulk, but they may as well correspond to various minor deviations from the model that are present in real data - we discuss this issue below.

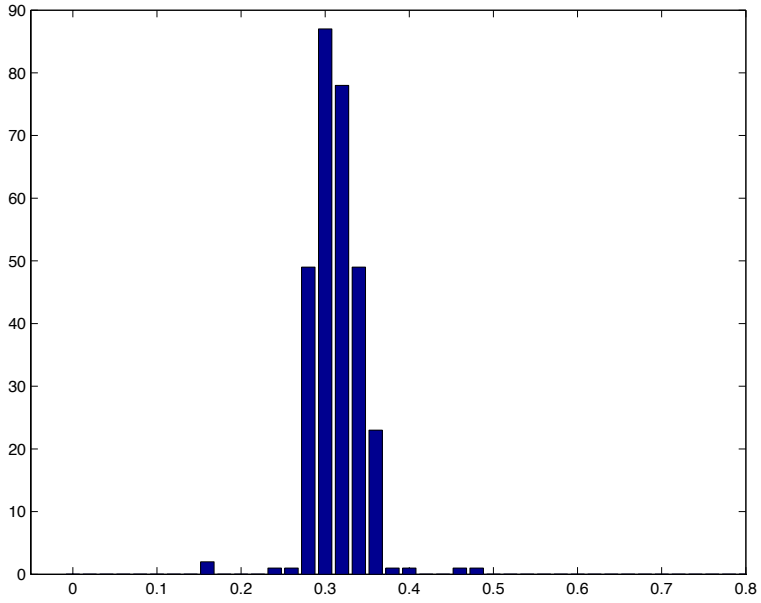


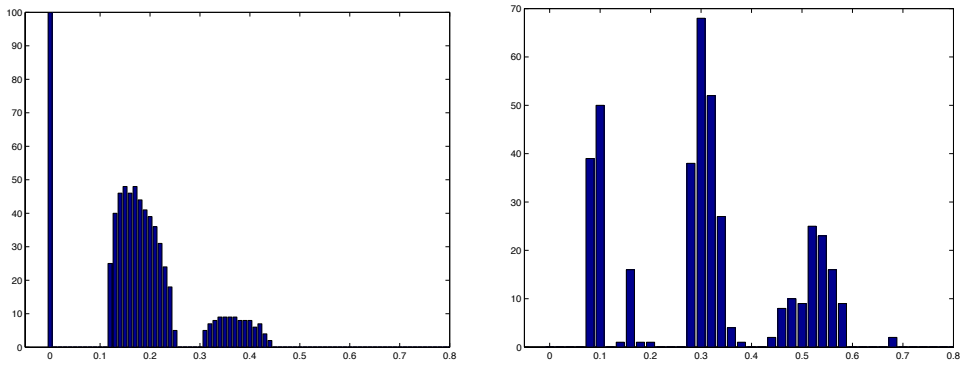
Figure 1: The bulk of the eigenvalues from PCA of Hapmap CEU, CHB, and YRI unrelated parents. The largest three eigenvalues that correspond to subpopulation structure are $\Lambda_1 = 102.0$, $\Lambda_2 = 14.55$, $\Lambda_3 = 7.37$ and are not shown.

3.4. Practical comments on using theory

Theorem 1 and the cutoff of 0.5 should be used in real data only after visual control for the shape of the bulk and for the separation from the bulk of the largest eigenvalues. Simulations indicate that when the theory is applicable the histogram of the bulk is located to the left of 0.5 (or 1 when $F = 1$) and its shape resembles the Marchenko-Pastur law ((Bai and Silverstein, 2010, Chapter 3)) of the same ratio N/M. For a typical large value of $N/M > 50$, this shape looks similar to a semi-ellipse.

The shape of the bulk is affected by relationships between the individuals. This is best illustrated when the offspring of trios are included in the analysis of the three populations HapMap dataset. Then the shape of the bulk does not follow Marchenko-Pastur law, and instead resembles a shape reproduced by

repeating a large number of individuals, see Figure 2.



$M = 700; N = 10000; N/M = 14.2$ $M = 405; N = 660847; N/M = 1631.7$

Figure 2: Including related individuals perturbs the expected shape of the bulk of the eigenvalues. *Left:* The bulk of the eigenvalues from PCA using data generated via binomial simulation, where 14% of the individuals have been repeated. *Right:* The bulk of the eigenvalues for PCA of three populations of HapMap (CEU, YRI, and CHB) including trios – 297 parents and their related 108 offspring. Both simulated data and empirical genotype data show that inclusion of related individuals results in a multi-modal distribution of the bulk of the eigenvalues, arising from the non-random correlations of individuals.

The shape of the bulk for the full HapMap data set seems to exhibit additional deviation from the expected shape, extends well beyond 0.5, and the distribution of the bulk might not be unimodal. These deviations cannot be attributed to linkage disequilibrium as they do not disappear after LD pruning. Instead, we expect these deviations from the expected shape correspond to complex substructure and relationships among individuals with insufficient power to be detected with so few individuals.

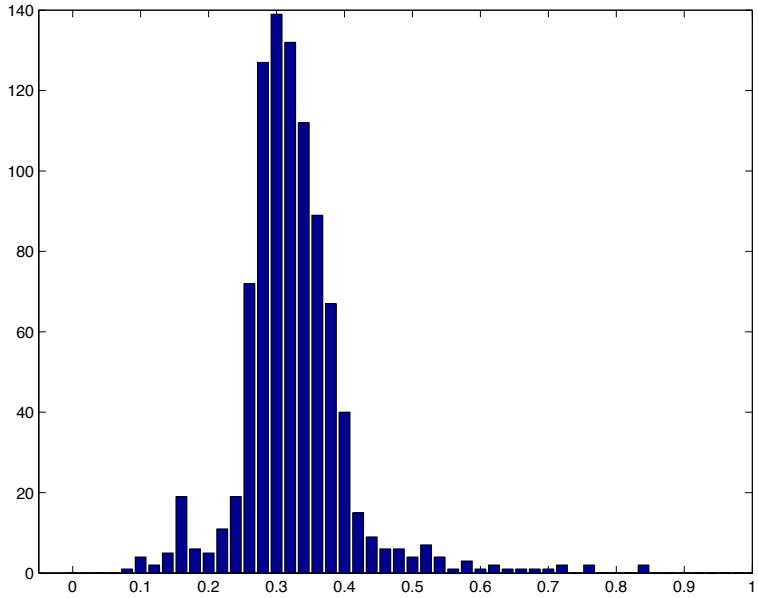


Figure 3: The bulk of the eigenvalues from PCA of all 11 populations in Hapmap unrelated parents. Nonautosomal markers with $M = 924$, $N = 422253$. The six largest eigenvalues $\Lambda_1 = 335.9$, $\Lambda_2 = 37.4$, $\Lambda_3 = 16.7$, $\Lambda_4 = 2.5$, $\Lambda_5 = 2.1$, $\Lambda_6 = 1.7$ are not shown

3.5. Conclusion

Eigenvalues of the uncentered covariance matrix $\mathbf{CC}'$ larger than the theoretical threshold (2.1), when combined with overall histogram of eigenvalues, are a consistent indicator of the presence of subpopulations in the data. We demonstrate in two proof-of-principle simulations that we are able to obtain evidence of population structure when the number of individuals is large enough. Our estimate is conservative, and we never encounter false positive evidence of population structure. That is, with independent markers, we never obtain evidence of more populations than present in the simulations. We encounter a loss of power (false negatives) in small simulated data sets. The accuracy of estimating the number of subpopulations K depends on the number of individuals M in the sample and is not improved significantly by increasing solely the number N of

available markers.

4. Theory

Our goal in this section is to point out aspects of population structure that could be responsible for the observed phenomenon that the set of eigenvalues of $\mathbf{CC}'$ splits into two groups: a small set of K large eigenvalues, and a large set of $M - K$ of small eigenvalues. Since $\mathbf{CC}'$ is a random matrix, we want this split to occur with overwhelming probability. This task requires a more detailed specification of the model. While the statements become more cumbersome, the gain is a clear indication of how different aspects of the model influence our ability to discover the subpopulation structure, and under what circumstances it may remain hidden.

We assume that our genetic markers are biallelic and that we have N polymorphic markers. We assume that we have diploid individuals from K subpopulations, and that each subpopulation r is described by its own allelic probability measures $\pi_{r,1}, \ldots, \pi_{r,N}$. That is, we assume that the data for the j-th marker of an individual from the r-th subpopulation take values $0, 1, 2$ with respective probabilities $\pi_{r,j}(0)$, $\pi_{r,j}(1)$, $\pi_{r,j}(2)$, which correspond to the frequencies of the genotypes in this population. Probabilities $\pi_{r,j}$ are a two-parameter family which we will parameterize by the half-mean: $p_r(j) = \pi_{r,j}(2) + \frac{1}{2}\pi_{r,j}(1)$ (allelic probabilities) and by the coefficients

$$1 - F_{r,j} = \frac{\pi_{r,j}(1)}{2p_r(j)(1 - p_r(j))},$$

so that the probabilities for the j-th marker of an individual from the r-th subpopulation take values $0, 1, 2$ with respective probabilities.

$$\pi_{r,j}(0) = (1-p_r(j))^2 + F_{r,j}\, p_r(j)(1-p_r(j)), \tag{4.1}$$

$$\pi_{r,j}(1) = 2p_r(j)(1-p_r(j))(1-F_{r,j}), \tag{4.2}$$

$$\pi_{r,j}(2) = p_r(j)^2 + F_{r,j}p_r(j)(1-p_r(j))). \tag{4.3}$$

We write

$$F = \sup_{r,j} F_{r,j} \tag{4.4}$$

for the least upper bound on these parameters. We always have the bound $0 \leq F \leq 1$, and $F = 0$ for a randomly mating population in Hardy-Weinberg equilibrium. The conservative value $F = 1$ is appropriate to use when the data depart significantly from Hardy-Weinberg equilibrium.

One possible interpretation of F is that it describes the least upper bound for the probability of mis-counting a heterozygote when processing the biological data; a probabilistic interpretation is that F is the so called ψ-mixing measure of dependence (Bradley, 2005) between the alleles, without attributing the causes of dependence. When $F_{r,j}$ does not depend on r, j another interpretation of their common value F is that this is an inbreeding coefficient (Wright, 1943).

The expected value of the law (4.1)-(4.3) is $2p_r(j)$ and the variance is $2(1 + F_{r,j})p_r(j)(1-p_r(j)) \leq (1+F)/2$.

In diffusion models, allelic probabilities $p_r(1), \ldots, p_r(N)$ are considered random and are then adequately described by their density function $\psi_r(x)$, $x \in [0,1]$, see e.g. (Kimura, 1964). We shall call $\psi_r(x)$ the site frequency spectrum for the r-th population.

Analysis of multi-population demographic models may be based on the joint site frequency spectra (Gutenkunst et al., 2009; Xie, 2010; Bustamante et al., 2001). For our analysis we do not really need the multi-variable joint frequency spectra; we only need to model each pair of populations r, s in terms of its pairwise

site frequency spectrum which is a probability measure $\Psi_{r,s}(dx, dy)$ on the unit square $[0,1] \times [0,1]$. In some situations, $\Psi_{r,s}(dx, dy)$ can be described by the corresponding density $\psi_{r,s}(x, y)$, and we adopt this for notational simplicity. Ref. (Gutenkunst et al., 2009) implements numerical approximation of joint frequency spectra for up to three simultaneous populations.

In practice, due to ascertainment bias the joint site frequency spectrum $\psi_{1,\ldots,K}(x_1, \ldots, x_k)$, or even the pairwise site frequency spectra $\psi_{r,s}(x, y)$ are often difficult to estimate from the genetic data. We describe population structure in SNP genotyping arrays by the ascertainment biased spectrum $\varphi_{r,s}(x, y)$ modeling each pair r, s of the populations. Consequently, we assume that allelic probabilities $p_r(1), \ldots, p_r(N)$ are random and are adequately described by their density function $\varphi_r(x)$, and that for the j-th locus each pair of allelic probabilities $(p_r(j), p_s(j))$ follow the same bivariate distribution with density $\varphi_{r,s}(x, y)$. We introduce the pairwise population moments:

$$m_{r,r} = \int_0^1 x^2 \varphi_r(x) dx, \tag{4.5}$$

and for $r \neq s$,

$$m_{r,s} = \int_0^1 \int_0^1 xy\, \varphi_{r,s}(x, y) dx dy. \tag{4.6}$$

The $K \times K$ array of deterministic numbers $m_{r,s}$ together with the relative subpopulation sizes are the parameters that enter our mathematical analysis.

Assumption 4.1. For any pair of subpopulations labeled by $r, s \in \{1, \ldots, K\}$, we assume that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} p_r(j) p_s(j) = m_{r,s} \tag{4.7}$$

In particular, we assume that $\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} p_r^2(j) = m_{r,r}$ exists.

A natural situation where the limit (4.7) exists is when the allelic probabilities have been sampled (independently, or with weak correlations) from a joint ascertainment biased site frequency spectrum $\varphi_{1,\ldots,K}(x_1, \ldots, x_k)$.

The allelic probabilities $p_r(1), \ldots, p_r(N)$ for the r-th subpopulation are of course unknown but represent the true underlying frequency of the alleles in the r-population and they are fixed when sampling the individuals from the population.

For $1 \leq r \leq K$ consider infinite sequences $\{p_r(j) : j \in \mathbb{N}\}$ of numbers in $[0, 1]$ such that the limits (4.7) exist. We also fix a sequence of constants $c_1, \ldots, c_K > 0$ that add up to 1.

Consider a $K \times K$ (deterministic) symmetric positive matrix $\mathbf{Q}$ with entries

$$[\mathbf{Q}]_{r,s} = \sqrt{c_r c_s} m_{r,s}, \tag{4.8}$$

where $m_{r,s}$ are given by (4.5) and (4.6). Next we assume that we have sequences $M_1(N), \ldots, M_K(N) \to \infty$ of integers such that with $M(N) = M_1(N) + \cdots + M_K(N)$, we have $M(N)/N \to c$ for some $c > 0$, and $M_r(N)/M(N) \to c_r > 0$ as $N \to \infty$. We assume that all the entries of matrix $\mathbf{C} = \mathbf{C}_N$ are independent, conditionally on $\{p_{r,j}\}$.

Since the singular values of $\mathbf{C}$ do not depend on the order of the rows, for mathematical analysis we assume that the individuals were arranged by population, so that $\mathbf{C}$ has block structure (4.9).

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \vdots \\ \mathbf{C}_K \end{bmatrix} \tag{4.9}$$

where $\mathbf{C}_r$ is the $M_r(N) \times N$ sub-matrix representing the data for the individuals from the r-th subpopulation. For $r = 1, \ldots, K$, $i = 1, \ldots, M_r$, $j = 1, \ldots, N$, we assume that the distribution of entry $[\mathbf{C}_r]_{i,j}$ is given by (4.1)-(4.3).

For the almost sure results we also need to assume that $\mathbf{C}$ comes from an infinite matrix. More specifically, we assume that conditionally on $\{p_{r,j}\}$, each of

the K blocks $\mathbf{C}_r$ arises as an upper-left $M_r \times N$ corner of an infinite matrix with independent entries that have distribution (4.1)–(4.3). We also need to make a technical assumption that the number of individuals from the r-th subpopulation increases with N, that is $M_r(N+1) \geq M_r(N)$.

Since the eigenvalues of $\mathbf{CC}'$ are large, it is more convenient to consider the normalized $M \times M$ sample covariance matrices

$$\mathbf{X}_N = \frac{1}{(\sqrt{M}+\sqrt{N})^2}\mathbf{CC}'. \tag{4.10}$$

Of course, we assume $N > M > K$.

Theorem 1. *Let $\lambda_1 \geq \lambda_2 \geq \lambda_K$ be the (deterministic) eigenvalues of matrix $\mathbf{Q}$ defined by* (4.8) *and let $\Lambda_1(N) \geq \Lambda_2(N) \geq \cdots \geq \Lambda_{M(N)}(N) \geq 0$ be the (random) eigenvalues of the $M(N) \times M(N)$ sample covariance matrix $\mathbf{X}_N$ from* (4.10). *Then, as $N \to \infty$, with probability one*

$$\Lambda_{K+1}(N) \leq (1+F)/2 \tag{4.11}$$

and

$$\left(\frac{1}{\sqrt{M(N)}}+\frac{1}{\sqrt{N}}\right)^2 [\Lambda_1(N), \Lambda_2(N), \ldots, \Lambda_K(N)] \to 4\,[\lambda_1, \lambda_2, \ldots, \lambda_K] \tag{4.12}$$

Remark 4.1. Under Hardy-Weinberg equilibrium $F = 0$, so (4.11) takes form

$$\Lambda_{K+1}(N) \leq 1/2. \tag{4.13}$$

However, usually the value of F is not known. In such cases, since $0 \leq F \leq 1$, while $\Lambda_K(N) \to \infty$ as $N \to \infty$ is much larger than 1, formula (4.11) may be replaced by

$$\Lambda_{K+1}(N) \leq 1. \tag{4.14}$$

When $\mathbf{Q}$ has full rank K formula (4.12) indicates that for large M, N the first K largest empirical eigenvalues of $\mathbf{X}_N$ are large and can be estimated from

the eigenvalues of $\mathbf{Q}$. Formula (4.11) shows that the remaining eigenvalues are relatively small, and are of the order $1/M$ smaller than the largest K eigenvalues (here we assume the standard setting in SNP data where $M < N$).

Remark 4.2. We have much less information for the case when $\mathbf{Q}$ has rank $K' < K$ with positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_{K'} > 0$ but $\lambda_{K'+1} = \cdots = \lambda_K = 0$. In this case the bulk is still concentrated below $1/2$, as (4.11) shows that $\Lambda_{K+1} \leq 1/2$. From (4.12) we deduce that the eigenvalues $\Lambda_1, \ldots, \Lambda_{K'}$ diverge to infinity. But we do not have any information about $\Lambda_{K'+1}, \ldots, \Lambda_K$ which in this case are only known to be of order smaller than $\frac{MN}{(\sqrt{M}+\sqrt{N})^2}$; we do not have any mathematical results about their relation to the"cutoff" $(1 + F)/2$.

4.1. Proof of Theorem 1

Let $P_k \in \mathbb{R}^N$ denote the column vector $[p_k(1), \ldots, p_k(N)]'$. Let $E_k \in \mathbb{R}^M$ be the column vector of ones at the rows corresponding to the k-th block of $\mathbf{C}$, i.e. with $[E_k]_r = 1$ if $M_1 + \cdots + M_{k-1} < r \leq M_1 + \cdots + M_k$ and 0 otherwise. Then $\mathbb{E}(\mathbf{C}) = 2\sum_{k=1}^K E_k P_k'$, and we write

$$\mathbf{C}_N = \mathbf{V} + 2\sum_{k=1}^K E_k P_k', \tag{4.15}$$

where $\mathbf{V}$ is an $M \times N$ matrix of centered independent uniformly bounded random variables. Note that $\mathbb{E}(\mathbf{C})$ factors as in (Engelhardt and Stephens, 2010, Eqtn. (1)), but we keep an additional term in analyzing (4.15).

Let $\lambda_1 \geq \cdots \geq \lambda_K \geq 0$ be all eigenvalues of $\mathbf{Q}$. (Recall that $N > M = M(N) > K$.)

Lemma 2. *Let $\sigma_1(N) \geq \sigma_2(N) \geq \cdots \geq \sigma_K(N) \geq 0$ be the singular values of $\sum_{k=1}^K E_k P_k'$. Then for $1 \leq j \leq K$, $\lim_{N\to\infty} \frac{\sigma_j^2(N)}{NM(N)} = \lambda_j$.*

Proof. Consider the sequence of $K \times K$ matrices $\mathbf{B}_N$ with entries

$$[\mathbf{B}_N]_{r,s} = \frac{\sqrt{M_r M_s}}{MN} \sum_{j=1}^N p_r(j) p_s(j). \tag{4.16}$$

(Recall that $M_r = M_r(N)$ is a function of N.) Since $\mathbf{B}_N \to \mathbf{Q}$ entrywise, its eigenvalues $\lambda_1(\mathbf{B}_N), \ldots, \lambda_K(\mathbf{B}_N)$ converge to $\lambda_1, \ldots, \lambda_K$. However, $\lambda_j(\mathbf{B}_N) = \frac{\sigma_j^2(N)}{NM}$ for all $j \in \{1, \ldots, K\}$. To see this, denote $U_k = \frac{1}{\sqrt{M_k}} E_k$. Then

$$\left(\sum_{k=1}^{K} E_k P_k'\right)\left(\sum_{k=1}^{K} E_k P_k'\right)' = NM \sum_{r,s=1}^{K} [\mathbf{B}_N]_{r,s} U_r U_s'$$

Let now $\vec{x} = [x_1, \ldots, x_K]'$ be an eigenvector corresponding to eigenvalue λ of $\mathbf{B}_N$. Then $\vec{y} = \sum_{k=1}^{K} x_k U_k \in \mathbb{R}^M$ is an eigenvector of $\sum_{r,s=1}^{K} [\mathbf{B}_N]_{r,s} U_r U_s'$ with the same λ. So $NM\lambda$ is the square of a singular value of $\sum_{k=1}^{K} E_k P_k'$. Note that orthogonal vectors $\vec{x}$ correspond to orthogonal $\vec{y}$, so this procedure exhausts the first K eigenvalues, even if they are repeated or 0; the remaining $M - K$ eigenvalues are zero . □

Lemma 3. *With probability one, as $M/N \to c > 0$,*

$$\limsup_{N\to\infty} \frac{1}{\sqrt{M} + \sqrt{N}} \|\mathbf{V}_N\| \le \sqrt{\frac{1+F}{2}}.$$

Proof. Recall that

$$\|\mathbf{V}\| = \sup\{\|\mathbf{V}x\|_{\mathbb{R}^M} : \|x\|_{\mathbb{R}^N} = 1\} = \sqrt{\lambda_1(\mathbf{V}\mathbf{V}')}$$

is a convex function of the entries of $\mathbf{V}$.

We apply a non-i.i.d. version of (Yin et al., 1988, Theorem 3.1), as extended in (Couillet et al., 2010, Theorem 3) to the matrix $\mathbf{B}_N = (\mathbf{V} + \mathbf{Z})/(\sqrt{M} + \sqrt{N})$, where $\mathbf{Z}$ is the matrix of independent two-valued random variables that compensate for the differences in the variances of entries of $\mathbf{V}$. That is, since

$$\mathbb{E}\left([\mathbf{V}]_{r,j}^2\right) = 2(1 + F_{r,j})p_{r,j}(1 - p_{r,j}) \le (1+F)/2, \tag{4.17}$$

we request that $\mathbb{E}([\mathbf{V}]_{r,j} + [\mathbf{Z}]_{r,j})^2 = (1+F)/2$.

Note that here we use a discrete distribution with just two values $[\mathbf{Z}]_{r,j} = \pm a_{r,j}$ such that $a_{r,j}^2 \le (1+F)/2 \le 1$. So assumption (2) of (Couillet et al., 2010,

Theorem 3) holds with a constant X, and assumption (3) of (Couillet et al., 2010, Theorem 3) holds with $\psi(x) = x^2$.

We also need to make sure that assumption (1) of (Couillet et al., 2010, Theorem 3) is satisfied, that is, we need to apply this theorem to a matrix that comes as an upper-left corner of an infinite matrix with independent entries.

To do so, we permute the rows of matrix $\mathbf{C}$ so that is arises as an $M(N) \times N$ sub-matrix from the infinite matrix which is described as follows. Recall that we assume that each matrix $\mathbf{C}_r$ comes as the upper-left corner from an infinite matrix of independent entries that correspond to population r. We now use these infinite matrices to construct a single infinite matrix, from which $\mathbf{C}$ arises by permuting the rows of the $M(n) \times N$ upper-left sub-matrix. We start with $N = 1$ and $M_1(1)$ (infinite) rows from population 1, followed by $M_2(1)$ rows from population 2, through $M_K(1)$ rows from population K. Then in the second "pass" we add $M_1(2) - M_1(1)$ rows from population 1, $M_2(2) - M_2(1)$ rows from population 2, etc. In the N-th pass, we add $M_1(N) - M_1(N-1)$ rows from population 1, $M_2(N) - M_2(N-1)$ rows from population 2, though $M_K(N) - M_K(N-1)$ rows from population K. Notice that since $M_r(N)$ are increasing in N, and $M(N) \to \infty$, this construction does not stop and gives an infinite matrix. It is clear, that after the N-th pass, the resulting $M(N) \times N$ sub-matrix will have exactly $M_r(N)$ rows from population r, so that its eigenvalues are the same as those of matrix $\mathbf{C}$.

Then, from (Couillet et al., 2010, Theorem 3) we infer that with probability one,

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}+\sqrt{M}} \|\mathbf{V}_N + \mathbf{Z}_N\| = \sqrt{\frac{1+F}{2}}.$$

Therefore, denoting by $\mu(d\mathbf{Z})$ the integral with respect to the law of the entire infinite sequence $Z_{i,j}$, by Jensen's inequality and Fatou's lemma we have

$$\begin{aligned}
\limsup_{N\to\infty} \|\mathbf{V}_N\|/(\sqrt{M}+\sqrt{N}) = \quad & \limsup_{N\to\infty} \| \int (\mathbf{V}_N + \mathbf{Z})\mu(d\mathbf{Z})\|/(\sqrt{M}+\sqrt{N}) \\
\leq \quad & \limsup_{N\to\infty} \int \|\mathbf{V}_N + \mathbf{Z}\|\mu(d\mathbf{Z})/(\sqrt{M}+\sqrt{N}) \\
\leq \quad & \int \left(\limsup_{N\to\infty} \|\mathbf{V}_N + \mathbf{Z}\|/\left(\sqrt{M}+\sqrt{N}\right)\right) \mu(d\mathbf{Z}) \\
= \quad & \sqrt{\frac{1+F}{2}}
\end{aligned}$$

with probability one. □

Proof of Theorem 1. This part of the proof is similar to (Silverstein, 1994). In (4.15), we consider $\mathbf{C}_N$ as a small perturbation of the finite rank matrix $2\sum_{k=1}^{K} E_k P_k'$.

Denote by $\tau_1(N) \geq \cdots \geq \tau_K(N)$ the largest(deterministic) singular values of

$$\frac{2}{\sqrt{N}+\sqrt{M}} \sum_{k=1}^{K} E_k P_k',$$

and set $\tau_j(N) = 0$ for $j > K$. Then it is known, see e.g. (Horn and Johnson, 1994, Theorem 3.3.16(c)), that the singular values $\sqrt{\Lambda_j}$ of $\frac{1}{\sqrt{N}+\sqrt{M}}\mathbf{C}$, written in decreasing order, differ by at most $\frac{1}{\sqrt{N}+\sqrt{M}}\|\mathbf{V}_N\|$ from the corresponding singular values $\tau_j(N)$, written in decreasing order.

Since $\tau_j(N) = 0$ for $j > K$, from Lemma 3 we get (4.11).

Lemma 2 shows that

$$\tau_j(N)(\sqrt{M}+\sqrt{N})/\sqrt{MN} = \tau_j(N)\left(\frac{1}{\sqrt{M}}+\frac{1}{\sqrt{N}}\right) \to 2\sqrt{\lambda_j}$$

for $1 \leq j \leq K$, so

$$\sqrt{\Lambda_j}\left(\frac{1}{\sqrt{M}}+\frac{1}{\sqrt{N}}\right)$$

has the same limit, and by taking squares of both sides we get (4.12). □

Acknowledgements

KB gratefully acknowledges support by the National Institutes of Health under Ruth L. Kirschstein National Research Service Award #5F32HG006411. The research of WB was partially supported by NSF grant #DMS-0904720.

References

Bai, Z., Silverstein, J., 2010. Spectral analysis of large dimensional random matrices (series: springer series in statistics) pod .

Bradley, R., 2005. Basic properties of strong mixing conditions. a survey and some open questions. Probability Surveys 2, 107–144.

Bustamante, C., Wakeley, J., Sawyer, S., Hartl, D., 2001. Directional selection and the site-frequency spectrum. Genetics 159, 1779.

Cavalli-Sforza, L., Menozzi, P., Piazza, A., 1993. Demic expansions and human evolution. Science 259, 639.

Cavalli-Sforza, L., Menozzi, P., Piazza, A., 1994. The history and geography of human genes. Princeton Univ Pr.

Couillet, R., Bai, Z., Debbah, M., Silverstein, J., 2010. Eigen-inference for energy estimation of multiple sources.

Engelhardt, B.E., Stephens, M., 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet 6.

Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587. Comparative Study.

Gao, H., Bryc, K., Bustamante, C., 2011. On identifying the optimal number of population clusters via the deviance information criterion. PloS one 6, e21014.

Gutenkunst, R., Hernandez, R., Williamson, S., Bustamante, C., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5, e1000695.

Horn, R., Johnson, C., 1994. Topics in matrix analysis. Cambridge Univ Pr.

Kimura, M., 1964. Diffusion models in population genetics. Journal of Applied Probability 1, 177–232.

McVean, G., 2009. A genealogical interpretation of principal components analysis. PLoS Genet 5.

Menozzi, P., Piazza, A., Cavalli-Sforza, L., 1978. Synthetic maps of human gene frequencies in Europeans. Science 201, 786.

Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M., Waeber, G., Vollenweider, P., Oksenberg, J.R., Hauser, S.L., Stirnadel, H.A., Kooner, J.S., Chambers, J.C., Jones, B., Mooser, V., Bustamante, C.D., Roses, A.D., Burns, D.K., Ehm, M.G., Lai, E.H., 2008. The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83, 347–358.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A., Auton, A., Indap, A., King, K., Bergmann, S., Nelson, M., et al., 2008. Genes mirror geography within Europe. Nature 456, 98–101.

Patterson, N., Price, A., Reich, D., 2006. Population structure and eigenanalysis. PLoS Genet 2, e190.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., De Bakker, P., Daly, M., et al., 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559–575.

Shriner, D., 2012. Improved eigenanalysis of discrete subpopulations and admixture using the minimum average partial test. Human Heredity 73, 73–83.

Silverstein, J.W., 1994. The spectral radii and norms of large-dimensional non-central random matrices. Comm. Statist. Stochastic Models 10, 525–532.

Wright, S., 1943. Isolation by distance. Genetics 28, 114.

Xie, X., 2010. The Site-Frequency Spectrum of Linked Sites. Bulletin of Mathematical Biology , 1–35.

Yin, Y., Bai, Z., Krishnaiah, P., 1988. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. Probability Theory and Related Fields 78, 509–521.

Zhu, X., Zhang, S., Zhao, H., Cooper, R., 2002. Association mapping, using a mixture model for complex traits. Genetic epidemiology 23, 181–196.